Report on Value Iteration

What is Value Iteration?

Value iteration is a special case of policy iteration, where the policy evaluation is stopped after one backup of each state. The current state can be obtained from the values in the backup state using the Bellman Equation.

Bellman Equation

$$V_{k+1}(s) \leftarrow \max_a \Sigma_{s'} T(s,a,s')(R(s,a,s') + \text{STEP} + \gamma V_k(s'))$$

Implementation

The details involved in the implementation of the Value Iteration is as given below:

Constants

The constants used in the question were stored in the parameter file `params.py`. Some individual constants are as mentioned below:

- **Step Cost (`STEP_COST`) =** -10
- **Discount Factor (`GAMMA`) =** 0.999
- **Bellman Error (`DELTA`) =** 0.001

> 💡 Our team number was 97 and 97%3 = 1, and hence the `STEP_COST` value was taken as -10 .

- **Shoot Damage (`FINAL_REWARD`) =** 25
- **Hit Damage (`HIT_DAMAGE`) =** 50
- **Cost of Successful Attack by MM (`MM_ATTACK_COST`) =** -40
- **Final Reward (`FINAL_REWARD`) =** 50

Reference Table (Positions and Corresponding Actions)

The actions corresponding to each position were stored in the form of a dicionary of dictionaries (3 - dimensions) within the parameter file `params.py`. While this is not a true transition table, it is a reference by which transition follows.

The stored values included possible outcomes of the actions, the transition probablities of each of the outcomes, the loss/gain of material or arrows during the action etc.

Storing States

States were stored in a 5-dimensional dictionary of dictionaries, taking the coresponding position, material, arrow, MM's state and MM's health (`< pos, mat, arrow, state, health >`) as the keys:

$$V\ [\text{pos}][\text{mat}][\text{arrow}][\text{state}][\text{health}]$$

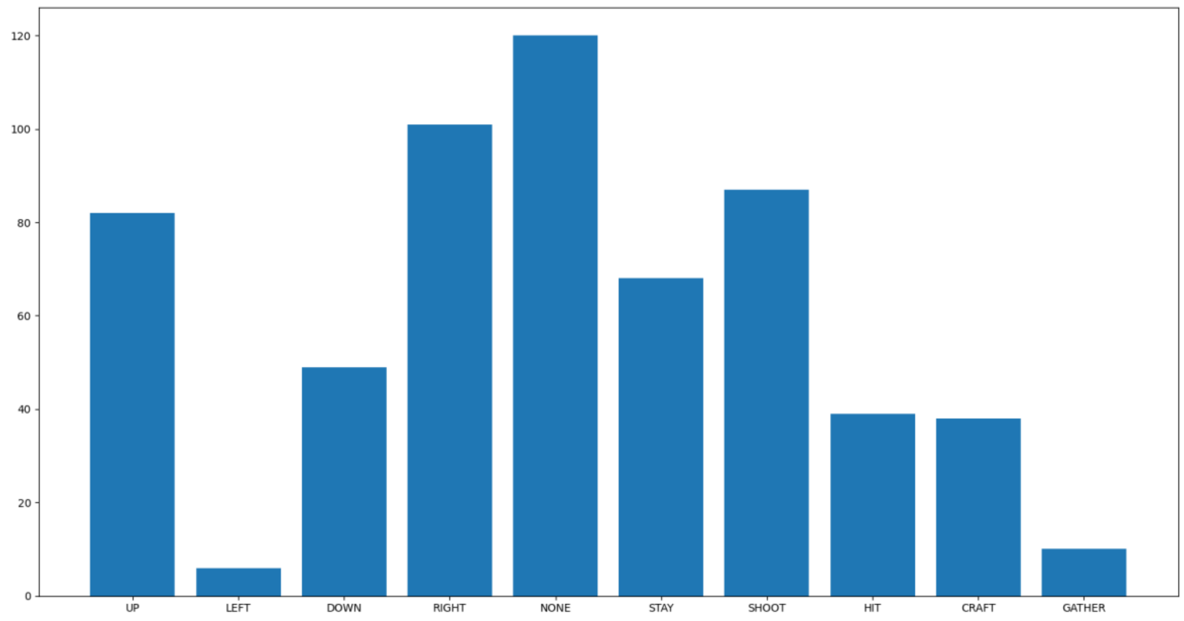
Task 1

Rate of Convergence

The Value Iteration converged in **118** iterations.

Action Frequencies

The action frequencies are as given the bar graph below:



Policies Obtained

On observation of Trace File:

The agent prefers to move to the East from current location and proceeds to `HIT` the monster from there.

Based on the Simulation with (W, 0, 0, D, 100) as Initial Point:

The agent moves `RIGHT` till it reaches East and then proceeds to `HIT` the monster from the East till termination.

Based on the Simulation with (C, 2, 0, R, 100) as Initial Point:

The agent moves `RIGHT` till it reaches East and then proceeds to `HIT` the monster from the East till termination.

Justification of the Policies Obtained

Due to the fact that step cost is so high, the agent prefers not to `GATHER` material or `CRAFT` at all. The move to East and performing `HIT` from there is justified by the high value of a successful `HIT`, the high value of `HIT_DAMAGE` and the low probablity of the transition of the monster to Ready state from the Dormant state.

Task 2

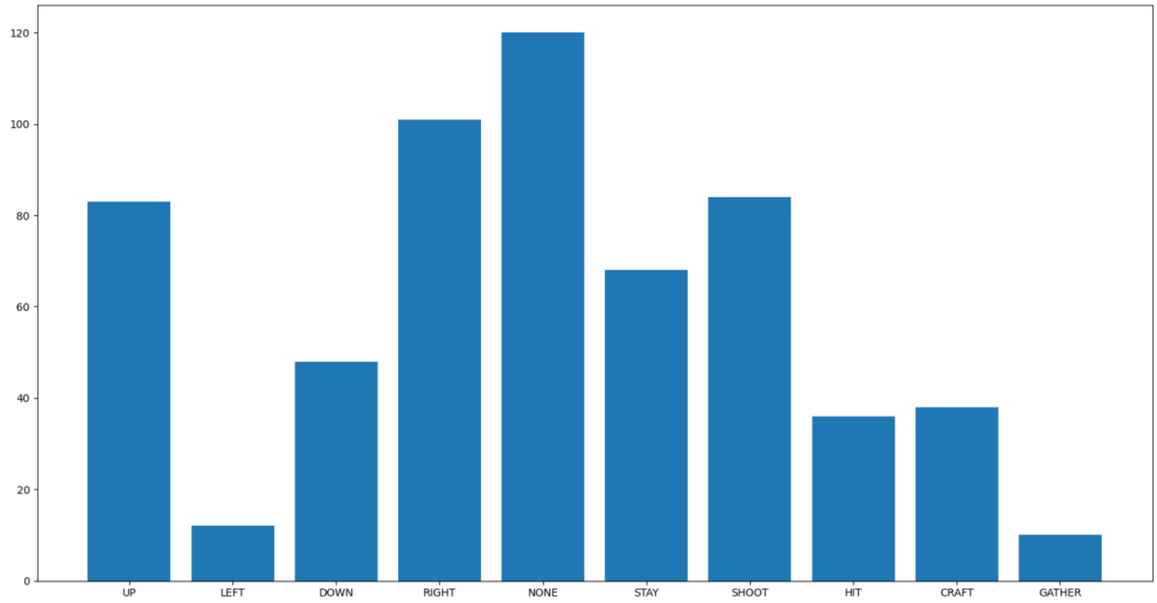
Case 1 → `LEFT` from East will go to the West

Rate of Convergence

The Value Iteration converged in **120** iterations.

Action Frequencies

The action frequencies are as given the bar graph below:



Policies Obtained

On observation of Trace File:

The agent prefers the action `CRAFT` while it has material and the action `SHOOT` while it has arrows. Else, it moves to the East from current location and proceeds to `HIT` the monster from there.

Based on the Simulation with (W, 0, 0, D, 100) as Initial Point:

The agent moves `RIGHT` till it reaches East and then proceeds to `HIT` the monster from the East till termination.

Based on the Simulation with (C, 2, 0, R, 100) as Initial Point:

The agent goes to North to `CRAFT` arrows and returns to center to `SHOOT` the monster. Once the arrows are depleted, the agent moves `RIGHT` till it reaches East and then proceeds to `HIT` the monster from the East till termination.

Justification of the Policies Obtained

Due to the fact that step cost is so high, the agent prefers not to `GATHER` material at all. The move to East and performing `HIT` from there is justified by the high value of a successful `HIT`, the high value of `HIT_DAMAGE` and the low probablity of the transition of the monster to Ready state from the Dormant state.

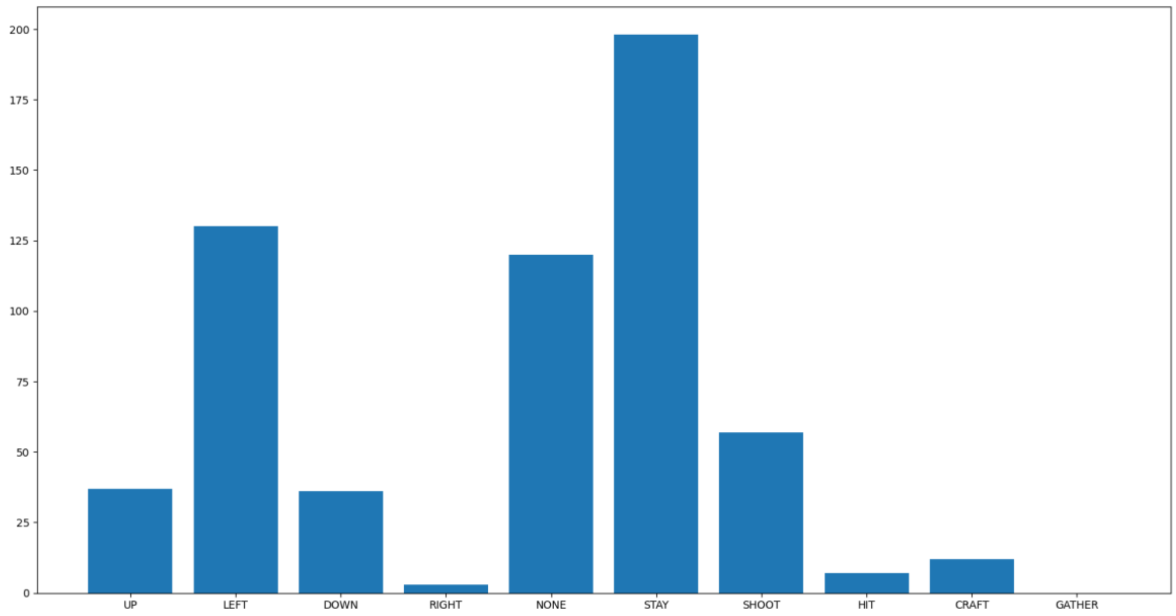
Case 2 → `STAY` action has 0 cost

Rate of Convergence

The Value Iteration converged in **57** iterations.

Action Frequencies

The action frequencies are as given the bar graph below:



Policies Obtained

On observation of Trace File:

In general, agent prefers to stay at its position. If on East or Center (places where the MM can attack), the agent prefers to move to the West from current location and proceeds to `STAY` the there.

Based on the Simulation with (W, 0, 0, D, 100) as Initial Point:

The agent performs the action `STAY` to remain on West. There is no termination observed in the simulation.

Based on the Simulation with (C, 2, 0, R, 100) as Initial Point:

The agent moves `LEFT` to go to the West. Once on the West, the agent performs the action `STAY` to remain on West. There is no termination observed in the simulation.

Justification of the Policies Obtained

Due to the fact that step cost is 0 for `STAY` and that it is so high for all other steps, the agent prefers to `STAY`. The move to the west is justified by the chance of attack at center by the monster would would result in -40.

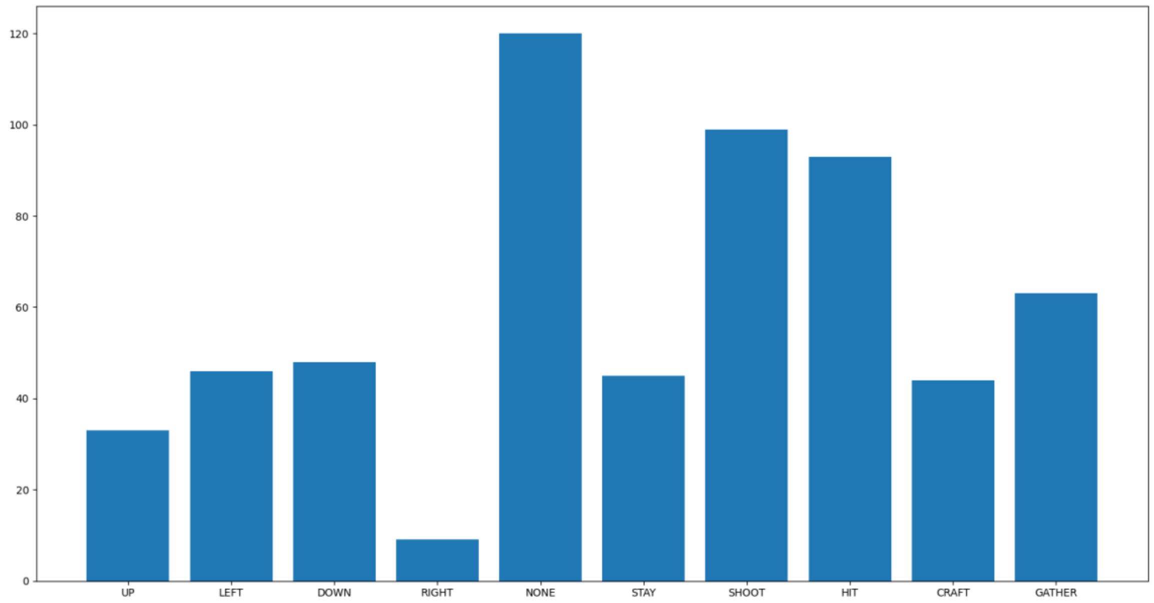
Case 2 →Discount Factor `GAMMA` is 0.25

Rate of Convergence

The Value Iteration converged in **8** iterations.

Action Frequencies

The action frequencies are as given the bar graph below:



Policies Obtained

On observation of Trace File:

In general, agent prefers to `SHOOT`, `HIT`, `GATHER` or `CRAFT` whenever possible. If the agent is on South, it prefers to `GATHER` even when the material is maximum.

Based on the Simulation with (W, 0, 0, D, 100) as Initial Point:

The agent performs the action `STAY` to remain on West. There is no termination observed in the simulation.

Based on the Simulation with (C, 2, 0, R, 100) as Initial Point:

The agent moves `DOWN` to go to the South. Once on the South, the agent performs the action `GATHER` to remain on South. There is no termination observed in the simulation.

Justification of the Policies Obtained

This is due to the very low `GAMMA` value, which has lead to excessive pruning. This causes the states to avoid going to the places where MM can attack, especially the East, at any cost. The `GATHER` on South can be explained by this as any other movement would lead to a small probablity of going to the East.